Projekt „DeutschDiachronDigital“

Technical Report 174 des Instituts für Informatik der Humboldt-Universität zu Berlin

Eine vergleichende Analyse von historischen und diachronen digitalen Korpora

Emil Kroymann, Sebastian Thiebes,

Anke Lüdeling, Ulf Leser

Humboldt-Universität zu Berlin
Institut für Informatik
Unter den Linden 6
D-10099 Berlin

Inhaltsverzeichnis

1. Einführung

Wie hat sich ein bestimmtes Wort/ein bestimmter Laut/eine bestimmte syntaktische Konstruktion verändert? Wie unterscheiden sich Geschäftsbriefe von Privatbriefen im 18. Jahrhundert? Wie unterscheiden sich Liebesbriefe des 18. Jahrhunderts von Liebesbriefen des 20. Jahrhunderts? Wie sehen überhaupt die ersten uns erhaltenen Liebesbriefe aus? Wann gab es die ersten Romane? Wie hat sich die Schrift entwickelt?

Zur Beantwortung dieser und ähnlicher Fragen werden – zusätzlich zu Buchausgaben – in den letzten Jahrzehnten alte Manuskripte und Drucke digitalisiert und über das Internet oder CDs zur Verfügung gestellt. Im vorliegenden Bericht werden solche sogenannten historischen Textkorpora (Korpora, die sich mit älteren Sprachstufen befassen) und diachronen Korpora (Korpora, die Texte aus mehreren Sprachstufen enthalten) vorgestellt und aufgrund einer Reihe von Kriterien bewertet.[1]

Ein Textkorpus[2] ist für uns eine annotierte, d. h., mit Zusatzinformationen versehene, elektronische Textsammlung für linguistische, computerlinguistische oder philologische Fragestellungen. Die Korpuslinguistik als eine Teildisziplin von Linguistik und Computerlinguistik beschäftigt sich mit der Zusammensetzung, Erstellung und Auswertung von Korpora. Obwohl die Korpuslinguistik seit ihren Anfängen auch historische Korpora erstellt (als erstes elektronisch verfügbares Korpus wurden in den 1940ern die Schriften von Thomas von Aquin digitalisiert, siehe Busa 1974, McEnery & Wilson 1998), sind die meisten heute verfügbaren Korpora synchrone Textsammlungen moderner Sprachen, und die meisten Verfahren zur Annotation und Auswertung sind auf diese Korpora zugeschnitten (siehe z. B. McEnery & Wilson 1998, Kennedy 1998, Manning & Schütze 1999, Klabunde et al. 2004).

Wir möchten die Bereiche Zusammensetzung, Erstellung und Auswertung von Korpora jeweils kurz generell erläutern, bevor wir auf spezielle Probleme von historischen und diachronen Korpora eingehen.[3]

Zusammensetzung von Korpora Die Zusammensetzung eines Korpus hängt von der wissenschaftlichen Fragestellung ab, die mit diesem Korpus untersucht werden soll. Man unterscheidet zwischen Referenzkorpora, die eine feste Größe und Zusammensetzung haben, und Monitorkorpora, die nach einem vorgegebenen Schema wachsen (z. B. indem jeweils täglich die neuste Ausgabe einer Tageszeitung hinzugefügt wird). Es gibt sehr spezielle Korpora, die Texte eines bestimmten Genres (z. B. nur Gedichte oder nur Emails) oder eines bestimmten Autors etc. enthalten, und große sogenannte ausgewogene oder repräsentative

---

[1] Diese Arbeit entstand als Vorbereitung für das Projekt „DeutschDiachronDigital" (DDD; http://www.deutschdiachrondigital.de/), das sich zum Ziel gesetzt hat, ein diachrones Korpus des Deutschen (Althochdeutsch bis frühes Neuhochdeutsch) zu erstellen.

[2] In der Korpuslinguistik unterscheidet man zwischen Textkorpora, die geschriebene Sprache enthalten, Sprachkorpora, die gesprochene Sprache enthalten und multimodalen Korpora, in denen verschiedene Kommunikationsmodi (gesprochene Sprache, Transkription, Gestik etc.) aligniert sind. Da für alte Sprachstufen nur Textdokumente vorhanden sind, verwenden wir hier die Begriffe Textkorpus und Korpus synonym. Die Sammlungen vieler Bibliotheken, die als reine Bilddigitalisate zur Verfügung gestellt werden, betrachten wir hier nicht.

[3] Wir können in diesem Report nicht die Geschichte und Entwicklung der historischen Korpuslinguistik nachzeichnen. Generell gilt, dass sich die historische Korpuslinguistik in einem Spannungsfeld zwischen den Möglichkeiten und Methoden der maschinellen Sprachverarbeitung (Natural Language Processing) und den Ansprüchen des sogenannten Humanistic Text Processing mit den Arbeiten zu elektronischen Editionen befindet. Für einen kurzen Abriss über die Anfänge der philologischen Korpuslinguistik siehe Zampolli (2004). Hockey (2003) bietet einen Überblick über digitale Ressourcen in den Philologien. Die Korpuslinguistik-Einführungen von McEnery & Wilson (1998) und Kennedy (1998) haben jeweils Kapitel über die Anfänge der Korpuslinguistik.

Korpora, die nach vorher festgelegten Parametern und Schlüsseln verschiedene Texte mischen (z. B. 10% gesprochene Sprache, 20% Privatbriefe, 15 % Zeitungstexte etc.). Die Begriffe 'ausgewogen' und 'repräsentativ' sind problematisch, da man ja die Grundgesamtheit nicht kennt (man müsste immer angeben, in Bezug auf welche Größe ein Korpus repräsentativ sein soll, siehe Biber 1993). Die genaue Aufteilung der Textsorten in großen Korpora unterscheidet sich dann auch je nach Fragestellung und Verfügbarkeit (vergl. zum Beispiel die Unterschiede beim British National Corpus BNC, http://www.natcorp.ox.ac.uk/, und der Textbasis für das Digitale Wörterbuch der Deutschen Sprache, http://www.dwds.de/).

Historische und diachrone Korpora sind im Prinzip immer Referenzkorpora (auch wenn sie in ihrer Entstehungsphase noch nicht genau festgelegt sind). Für diachrone Untersuchungen – seien dies Untersuchungen zum Sprachwandel oder auch philologische Untersuchungen – ist eine Vergleichbarkeit über verschiedene Sprachstufen hinweg nötig; im Idealfall sollte sich also nur der Parameter Zeit unterscheiden, während alle anderen Parameter wie Textsorte, Formalisierungsgrad etc. gleich bleiben. Das ist natürlich bei älteren Sprachstufen noch schwerer möglich als bei modernen, zum einen, weil sich viele Textsorten (wie z. B. Romane oder Tageszeitungsberichte) erst später entwickelt haben und andere (wie z. B. Evangelienharmonien) verschwunden sind und zum anderen, weil viel weniger Material erhalten ist. Man bekommt also immer nur in Teilbereichen Kontinuität. Wenn man eine Matrix aller relevanten Parameter aufstellt (z. B. Sprachstufe, Textsorte und Dialekt), bleiben bei historischen Korpora zwangsläufig einige Zellen leer (dazu siehe z. B. die Diskussionen in Rissanen et al. (1993)).

Eine wichtige Voraussetzung für diachrone Studien genauso wie für vergleichende Studien innerhalb einer Sprachstufe ist eine detaillierte Annotation der Texte in einer sogenannten Headerstruktur. Hier werden Angaben zu Textsorte, Autor, intendiertem Rezipientenkreis, Dialekt etc. nach vorher festgelegten Kriterien vermerkt. Daraus kann man sich dann je nach Bedarf Teilkorpora zusammenstellen (z. B. alle Texte zwischen 1600 und 1650 oder alle von Frauen geschriebenen Texte). In den Headerstrukturen sind im Idealfall auch Informationen zur Art der Verarbeitung (welche Vorlage wurde verwendet, wie wurde digitalisiert, wie diplomatisch wurde gearbeitet, wie wurde korrigiert etc.) angegeben.

Zur Annotation von Headerstrukturen haben sich einige internationale Standards herausgebildet. Während früher zum Teil nach dem COCOA-Standard annotiert wurde, haben sich heute weitgehend der Corpus Encoding Standard (CES) und die Empfehlungen der Text Encoding Initiative (TEI, http://www.tei-c.org/, die XML Version des TEI-konformen Corpus Encoding Standard CES befindet sich unter http://www.xces.org/) durchgesetzt. Für historische Texte braucht man zum Teil andere Informationen (z. B. Informationen zum Schrifttyp eines Manuskripts) als für moderne – hier ist die TEI dabei, Vorschläge zu entwickeln.

Erstellung von Korpora Wenn die Zusammenstellung eines Korpus entschieden ist, muss man die Texte digitalisieren oder bereits elektronisch vorhandene Texte in ein einheitliches Format bringen. Dann werden die Texte tokenisiert und positionell und strukturell annotiert. Die Tokenisierung teilt das Korpus in einzelne sogenannte 'Token' oder Textwörter – für Sprachen wie das Deutsche und Englische sind dies meist graphemische Wörter, d. h. Zeichenketten zwischen Satz- und Leerzeichen (dies ist in vielen Fällen problematisch, vgl. Fälle wie *New York* in denen zwei Token zusammen einen Namen ergeben einerseits und Fälle wie *beim* oder *siehste*, in denen ein Token in zwei lexikalische Wörter aufgeteilt werden müsste). Strukturelle Annotation markiert einerseits die graphische (wie z. B. Zeilen oder Seiten) und andererseits die logische Struktur des Textes (Absätze, Kapitel). Positionelle Annotationen weisen einzelnen Token, die an einer bestimmten Korpusposition stehen,

Informationen zu, typischerweise Wortart, Lemma oder flexionsmorphologische Informationen (im Prinzip kann man hier alle gewünschten Informationen einfügen; einige Korpora sind z. B. syntaktisch annotiert). Dazu wird für jede Annotationsebene ein Tagset entwickelt, in dem die möglichen Werte (also zum Beispiel die möglichen Wortarten) angegeben sind. Bei modernen Sprachstufen erfolgt die Annotation meist automatisch – dafür sind verschiedene regelbasierte oder statistische Verfahren entwickelt worden (ein Überblick über die Vorverarbeitung von Korpora findet sich in Evert & Fitschen 2004, siehe auch Manning & Schütze 1999). Die Fehlerraten bei automatischer Annotion unterscheiden sich je nach Annotationsebene – die besten Wortarttagger für das Deutsche erreichen eine Korrektheit auf Zeitungstext zwischen 95 und 98%.

Für Korpora älterer Sprachstufen müssen schon in der Phase der Digitalisierung viele Entscheidungen getroffen werden. Zunächst muss geklärt werden, ob von einem Originaltext (also Handschrift oder früher Druck) oder von einer Edition digitalisiert wird. Dann muss der Grad der Diplomatizität (d. h. die Nähe zum Text) festgelegt werden – dabei sind Entscheidungen über Sonderzeichen, paläographische Besonderheiten, Kodierung von Unsicherheiten etc. zu treffen. Dabei ist es wichtig, dass Angaben zur Digitalisierung in der Headerstruktur angegeben werden, so dass eine Nutzerin feststellen kann, ob ein Text für eine Forschungsfrage geeignet ist.

Auch die Annotation von Texten älterer Sprachstufen ist in einigen Bereichen schwieriger – das liegt zum einen an der niedrigen Standardisierung und andererseits an den fehlenden Ressourcen wie maschinenlesbaren Lexika etc. Bei Handschriften ist schon die Tokenisierung ein Problem, da graphisches Wort und lexikalisches Wort oft nicht übereinstimmen. Die Tagsets, die für moderne Sprachstufen entwickelt wurden, sind nicht einfach auf ältere Sprachstufen übertragbar – besondere Schwierigkeiten ergeben sich bei diachronen Korpora, da sich sowohl die Form als auch die Funktion von Wortarten über die Zeit ändern kann. Daher sind nur wenige Korpora älterer Sprachstufen bisher positionell annotiert. Ein weiteres Problem ist die Fehlerrate: in der historischen Linguistik sind Fehlerraten von 5% (wenn sie denn überhaupt erreichbar wären) nicht akzeptabel (denn immerhin ist da noch jedes 20. Wort falsch). Daher werden ältere Sprachstufen manuell oder semi-automatisch annotiert – das bedeutet, dass nur relativ kleine Korpora überhaupt annotiert werden können.

Auswertung von Korpora Textdaten kann man quantitativ und qualitativ auswerten. Die erste Auswertungsstufe ist die Volltextsuche – dabei unterscheidet sich ein elektronisch vorliegendes Korpus zunächst einmal nicht prinzipiell von nicht elektronisch vorliegenden Texten. Die ersten Korpusabfragesprachen, die entwickelt wurden, produzierten sogenannten kwic-Konkordanzen (für keyword-in-context), die nach den vorher auf Papier vorliegenden Konkordanzen ('Zettelsammlungen') modelliert waren.

Allerdings können elektronisch vorliegende Korpora systematisch und schnell durchsucht und umgeordnet werden – die Ergebnisse werden dadurch reproduzierbar, und alternative Hypothesen können schnell überprüft werden. Durch Headerinformationen können jeweils geeignete Textsammlungen zusammengestellt werden. Annotierte Ebenen können mit dem Text durchsucht werden. Zusätzlich zur Volltextsuche sind in der Korpuslinguistik mächtige Abfragesprachen entwickelt worden (zum Beispiel die Abfragesprache CQP, die die volle Mächtigkeit regulärer Ausdrücke hat, siehe http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/CQPUserManual/HTML/). Es gibt inzwischen auch Abfragesprachen für Bäume und Graphen (z. B. TigerSearch http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERSearch/). Für die lexikographische Arbeit gibt es inzwischen eine Reihe von Werkzeugen, die die Kookkurrenzstärke von zwei Einträgen messen. Andere arbeiten auf syntaktischen Annotationen und errechnen für ein Wort

'typische' (d. h. häufige oder statistisch signifikante) Subkategorisierungs- und Modifikationsangaben.

In vielen historischen Korpora werden die aktuellen Möglichkeiten bisher nicht voll genutzt (Ausnahmen sind die Helsinki-Corpus-Familie, die die Tools nutzt, die für die Penn Treebank entwickelt wurden und die Texte im TITUS Korpus) – zum Teil ist nur eine Volltextsuche möglich. Nur einige Korpora haben überhaupt Headerinformationen. Die Korpora, die sich nicht an die COCOA oder TEI Standards halten, sind dann überhaupt nicht vergleichbar.

Quantitative Auswertungsmöglichkeiten wie Kollokationsanalyse stehen nur in wenigen Fällen zur Verfügung.

1.1. Übersicht über historische und diachrone Korpora

Im Folgenden stellen wir eine Übersicht über vorhandene historische und diachrone Korpora vor.[4] Die historische Korpuslinguistik hat sich vor allem anhand des Englischen entwickelt – viele Techniken und Standards wurden hier erprobt (zum Beispiel die syntaktische Annotation historischer Texte). Deshalb haben wir uns auf deutsche und englische Korpora konzentriert und in anderen Sprachen nur große Projekte ausgewählt. Wir haben keine Projekte aufgenommen, die reine Bilddigitalisierungen von Manuskripten oder frühen Drucken herstellen, da es uns um linguistische Auswertungsmöglichkeiten geht. Außerdem erwähnen wir hier keine Projekte, die nur einen einzelnen Text digitalisiert haben.[5]

Die Korpusprojekte, die wir betrachtet haben, sind sehr unterschiedlich, sowohl was die Zielsetzung angeht, mit der sie ins Leben gerufen wurden, als auch in den verwendeten Methoden zur Erstellung und Vorhaltung des Korpus. Zum ersten gibt es Korpora, die den wissenschaftlichen Anspruch haben, eine gewisse Sprache aus einer bestimmten Zeit repräsentativ darzustellen oder verschiedene Sprachregionen möglichst ausgewogen zu erfassen, und zum zweiten gibt es Korpora, die aus enger gefassten wissenschaftlichen Interessen entstanden sind und zum dritten sogenannte opportunistische Textsammlungen, die ohne besonderen Anspruch an Ausgewogenheit oder Repräsentativität erstellt wurden. Die Methoden und Formate, in denen die Daten aufbereitet und bereitgestellt werden, sind selten standardisiert. Es überwiegen individuelle Lösungen, und leider sind diese oft nicht genauer beschrieben. Eine Ausnahme bildet hier das TEI-konforme SGML, an das sich einige der Korpusarchitekturen als einen minimalen Datenrepräsentationsstandard halten.

Für den Vergleich haben wir die folgenden Untersuchungskriterien gewählt.

---

[4] Wir beschreiben hier die uns zugänglichen Korpora und Digitalisierungsprojekte. Falls wir ein Projekt übersehen haben, würden wir uns über eine Benachrichtigung freuen.

[5] Dazu gehören auch kommerzielle Editionsprojekte. Korpora, die überhaupt nicht öffentlich zugänglich oder beschrieben sind, werden ebenfalls nicht erwähnt (davon gibt es wahrscheinlich viele; wir kennen einige – wie das Mainzer Zeitungskorpus – aus Veröffentlichungen, können aber nichts über Zusammensetzung, Annotation etc. herausfinden).

1.2. Erläuterung der Untersuchungskriterien

1.2.1. Textauswahl & Digitalisierungsverfahren

Textauswahl: Die Textauswahl in den betrachteten Korpora ist sehr unterschiedlich: einige enthalten opportunistisch gesammelte Texte, andere versuchen, Ausgewogenheit nach bestimmten Kriterien (Dialekte, Sprachstufen, Genres etc.) herzustellen, wieder andere sammeln sehr spezifisch Texte nach einem einzigen Parameter (wie z. B. Texte eines Autors oder eines Genres).

Größe: Auch hier gibt es große Unterschiede, die Umfänge reichen von einigen Tausend Wörtern bis zu mehreren Millionen Wörtern. Gerade Korpora, die ältere Zeitepochen abdecken, sind aufgrund der Quellenlage oft relativ klein.

Textgrundlage, Diplomatizität, Qualitätssicherung: Als erstes werden die verwendeten Techniken und Werkzeuge, mit deren Hilfe die Texte des jeweiligen Korpus digitalisiert wurden, aufgezählt und gegebenenfalls erläutert. Ein wichtiges Kriterium ist hier, insbesondere bei Handschriften, der Grad der Diplomatizität. Leider gibt es hierzu keine einheitlichen Standards, und es wird in den Dokumentationen der Korpora selten erwähnt, nach welchen Kriterien transkribiert wurde: oft weiß man nicht, wie mit Sonderzeichen, Abkürzungen, Marginalien etc. aus alten Handschriften umgegangen wurde und inwieweit normalisiert wurde.

Viele Korpora benutzen auch oder ausschließlich Materialien, die aus schon vorhandenen Editionen stammen. Auch hier können unterschiedliche Grade an Diplomatizität festgestellt werden (wie genau bildet der E-Text die Edition ab, werden Fehler in der Edition korrigiert, werden strukturelle Informationen wie Zeilenenden oder Versnummerierung mitkodiert etc?). Hängt die Qualität der Texte eines aus Originalquellen erstellten Korpus hauptsächlich von den Methoden der Erfassung und deren Genauigkeit ab, so ist die Übernahme von Editionstexten fast ausschließlich von der Qualität der Edition abhängig, die eventuell durch eine ungenaue Erfassung noch negativ beeinflusst werden kann.

1.2.2. Annotation

Wesentlich für den Vergleich der Korpora ist der Umfang der Annotationen, d. h. die Anzahl der verschiedenen Ebenen, die ausgezeichnet wurden. Von Interesse sind außerdem wiederum die Techniken und die verwendeten Werkzeuge, die für die Annotation verwendet wurden.

Wir unterscheiden dabei wie oben erläutert zwischen struktureller Annotation, positioneller Annotation und Headerinformationen. Strukturelle Annotation bezieht sich auf die graphische (Zeile, Seite) oder die logische (Satz, Absatz, Kapitel) Struktur eines Textes. Man muss dabei darauf achten, ob die graphischen Annotationen sich auf eine Edition beziehen (und dabei z. B. auch Zeilennummern aufführen) oder auf den Originaltext. Positionelle Annotationen bezeichnen Informationen, die sich auf eine bestimmte Korpusposition (in der Praxis fast immer auf ein Token) beziehen. Typische positionelle Annotationen sind Lemmaangaben, Normalisierung, Wortart oder flexionsmorphologische Angaben. In manchen Korpora sind den Texten Metainformationen – die sogenannten Headerinformationen - mitgegeben, in denen zum Beispiel Herkunft, Alter, Autor, Genre oder Dialekt eines Textes kodiert sind.

Da auch bei der Annotation kein einheitlicher Standard bzgl. des Tag-Sets besteht, ist es wesentlich den verwendeten Standard zu benennen bzw. zu beschreiben. In der

Korpuslinguistik sind gemeinsame Annotationsstandards vorgeschlagen worden (so zum Beispiel die Annotationsrichtlinien der Text Encoding Initiative http://www.tei-c.org/ oder der Corpus Encoding Standard http://www.xml-ces.org/). Wir haben bei der Auswertung angegeben, wenn ein solcher Standard beachtet wurde.

1.2.3. Suchmöglichkeiten

Hier werden die Suchmöglichkeiten beschrieben, die mit dem Korpus mitgeliefert werden.

Viele der Korpora sind öffentlich über das World Wide Web (WWW) zugänglich. Bei diesen gibt es oft die Möglichkeit das Korpus mittels einer speziellen Suchmaske zu durchsuchen. Aber auch bei Korpora, die als eine oder mehrere Dateien offline verfügbar oder bestellbar sind, gibt es oft die Möglichkeit, mit speziellen Programmen nach Textstellen, Annotationen o. ä. zu suchen. Dabei wird angegeben, ob eine reine Volltextsuche oder auch mächtigere Möglichkeiten zur Suche (wildcards, reguläre Ausdrücke, ...) vorhanden sind.

Die Suchergebnisse variieren in Art und Umfang. So ist bei manchen lediglich die Zeilennummer des Ergebnisses angegeben, während bei anderen ein Ausschnitt des Textes um die Fundstelle als Ergebnis angezeigt wird.

Zu manchen Korpora werden externe Wissensquellen wie Lexika etc. zur Verfügung gestellt.

1.2.4. Verfügbarkeit

Wie schon erwähnt, sind viele der Korpora öffentlich über das WWW zugänglich. Manche unterliegen aber Restriktionen, so dass sie z. B. nur für die universitäre Lehre oder nach der Zahlung einer Gebühr nutzbar sind. So sind auch teilweise die enthaltenen Texte durch Copyright geschützt, so dass sie nicht öffentlich zugänglich gemacht werden dürfen und nur Forschungspartner des jeweiligen Projektes zur Verfügung stehen.

1.2.5. Projektinformation & Kontakt

In diesem Abschnitt werden Informationen über das Projekt oder die Organisation, die das Korpus erstellt hat, angegeben. Wenn möglich, wird eine Kontaktadresse genannt. Außerdem wird hier angegeben, ob ein Korpus aktuell noch gepflegt wird.

1.2.6. Literatur

Ein weiteres wichtiges Qualitätskriterium für bestehende Korpora ist der Grad der Dokumentation. Hier werden sowohl online Dokumentationen (Handhabung, verwendete Software, Standards) als auch Literatur zu den Korpora aufgelistet.

1.2.7. Spezielles

Jegliche Information, die sich nicht durch die oben genannten Punkte klassifizieren lässt.

2. Historische Korpora des Deutschen

2.1. Bonner Frühneuhochdeutsch Korpus

Textauswahl & Digitalisierungsverfahren

Sowohl Handschriften als auch Editionen sind in das Korpus eingeflossen. Insgesamt sind es 40 Texte mit je ca. 400 Zeichen. Die Texte stammen aus zehn deutschen Sprachräumen und je vier Zeitabschnitten (1350 - 1400, 1450 - 1500, 1550 - 1600 und 1650 -1700).

Annotation

Wortformen wurden manuell mit einem eigenen Tag-Set ausgezeichnet. Verben, Nomen und Adjektive wurden mit Informationen zu Wortart, Flexionsklasse und grammatischen Merkmalen versehen (Finitheit, Person, Numerus, Tempus und Modus für Verben, Genus und Kasus für Nomen und Adjektive). Des Weiteren sind Quellenangaben zu den annotierten Texten vorhanden. Die Texte sind in drei Formaten vorhanden: Als XML oder HTML-kodierte Fassung oder in der Originalfassung, die aus der Übertragung von Lochkarten entstand und mit dem Zeichensatz der MS-DOS Codepage 437 gespeichert wurde.

Suchmöglichkeiten

Es gibt keine Suchmöglichkeiten.

Verfügbarkeit

Das Korpus ist per Download frei verfügbar.

Kontakt

Projektwebseite: http://www.ikp.uni-bonn.de/dt/forsch/fnhd/

Verantwortlich: Winfried Lenders <Lenders@uni-bonn.de>, Hans-Christian Schmitz <hcs@ikp.uni-bonn.de>

Literatur

Lender, Winfried; Klaus-Peter Wegera (Hrsg.): *Maschinelle Auswertung sprachhistorischer Quellen. Ein Bericht zur computerunterstützten Analyse der Flexionsmorphologie des Frühneuhochdeutschen*. Tübingen: Niemeyer 1982 (= Sprache und Information Bd. 3).

2.2. Bochumer Mittelhochdeutsch Korpus

Textauswahl & Digitalisierungsverfahren

Korpus von Lyrik und Prosatexten aus dem Zeitraum 1070-1350. Es sind Texte aus den Sprachräumen Oberdeutsch, Bairisch, Bairisch-Alemannisch, Alemannisch Westmitteldeutsch, Hessisch-Thüringisch, Schwäbisch Mittelfränkisch, Rheinfränkisch, Ostmittelfränkisch, Ostfränkisch enthalten. Die Texte sind nach Editionen digitalisiert.

Keine Angaben dazu in welcher Form das Korpus vorliegt (d. h., u. U. nicht digital).

Der Umfang beträgt 105 Texte.

Annotation

Keine Angaben zur Annotation

Suchmöglichkeiten

Keine Angaben zu den Suchmöglichkeiten

Verfügbarkeit

Keine Angaben zur Verfügbarkeit

Projektinformation & Kontakt

Webseite: www.ruhr-uni-bochum.de/wegera/archiv

Verantwortlich: Klaus-Peter Wegera <klaus-peter.wegera@ruhr.uni-bochum.de>

2.3. Mittelhochdeutsche Wörterbücher und Digitales Mittelhochdeutsches Textarchiv

Textauswahl & Digitalisierungsverfahren

Die Zielsetzung des DFG-Projekts besteht zum einen darin, das Mittelhochdeutsche Wörterbuch von Benecke/Müller/Zarncke (BMZ), das Mittelhochdeutsche Handwörterbuch von Lexer (Lexer) und das Findebuch zum mittelhochdeutschen Wortschatz (Findebuch) zu digitalisieren.

Die Digitalisierung der Wörterbücher erfolgte durch chinesische Dienstleister durch zweifache Eingabe von Hand.

Zum anderen sollen die Einträge der Wörterbücher mit Belegstellen in einem digitalen Belegarchiv verknüpft werden. Dazu soll ein Korpus von philologisch gesicherten Texten erstellt werden. Das Korpus soll zunächst 148 mittelhochdeutsche Texte und Textsammlungen, aus Editionen übernommen, enthalten.

Die Texte sind im TEI XML-Format codiert. Nicht darstellbare Sonderzeichen werden mit Hilfe von XML-Entities codiert.

Annotation

Die Texte im Korpus sind mit verschiedenen Metainformationen zum Text und zur Edition annotiert. Unter anderem Autor, Titel, Gliederung des Textes (Kapitel, Verse), Anmerkungen des Editors, Seite und Zeile.

Suchmöglichkeiten

Volltextsuche über den gesamten Wörterbuchinhalt

In der CD-ROM-Version ist zusätzlich die Suche nach grammatischen Angaben, Belegstellen, einzelnen Siglen und nach Textsorten zusammengefassten Siglen möglich.

Die Suche im Korpus wird durch ein Webinterface ermöglicht.

Die Suchmöglichkeiten sind zum einen ein so genannter Textbrowser. Der Textbrowser stellt einen Text seitenweise wie in der Textausgabe dar. Er ermöglicht außerdem den Zugriff auf die Metainformationen zum Text.

Eine weitere Suchmöglichkeit ist die Volltextsuche. Über diese kann man nach Wortformen die im Text vorkommen suchen. Die Vorkommen können einfach aufgelistet werden und als Konkordanz oder im Kontext angezeigt werden.

Zusätzlich gibt es die Möglichkeit einzelne Texte im TEI XML-Format herunterzuladen.

Verfügbarkeit

Eine eingeschränkte Version ist im Internet frei zugänglich. Die CD-ROM Version ist über den Verlag S. Hirzel, Stuttgart erhältlich. Eine Linux Version der CD-ROM vertreibt die Uni Trier.

Projektinformation & Kontakt

Darstellung des Projekts: http://gaer27.uni-trier.de/MWV-online/MWV-online.html

Darstellung des Korpus: http://www.mhgta.uni-trier.de/Demo

Verantwortlich:

Arbeitsstelle Mittelhochdeutsches Wörterbuch, http://www.mhdwb.uni-trier.de/

Leitung: Ralf Plate <plate@uni-trier.de>

2.4. Thesaurus Indogermanischer Text- und Sprachmaterialien(TITUS)

Textauswahl & Digitalisierungsverfahren

TITUS ist eher ein Archiv als ein Korpus. Es enthält Texte aus dem gesamten indogermanischen Sprachraum aus verschiedenen Zeiträumen. Die genaue Aufstellung der Texte ist unter http://titus.uni-frankfurt.de/texte/texte.htm zugänglich (das Archiv wird laufend erweitert).

Die Texte in TITUS sind zumeist Digitalisierungen bereits in Druckform ediert vorliegender Textmaterialien, wobei die Texte wie in den gedruckten Editionen eingegeben werden. Der zweite Schritt besteht in der Kollationierung und Überarbeitung bezogen auf die Originalhandschriften, wobei aus Qualitätsgründen, aber auch zum Zwecke der Konservierung, der Umweg über die Erstellung von Farbdias gewählt wurde, die dann mit einem hochauflösenden Dia-Scanner eingelesen werden. Geplant ist eine zumindest teilweise Codierung in TEI-XML.

Annotation

Es sind Informationen zu Sprache bzw. Sprachstufe, Text, Autor, Kapitel, Absatz, Vers, Seite, Zeile und Lemmainformationen erfasst. Einige Texte sind mit WordCruncher aufbereitet.

Suchmöglichkeiten

Über ein Webinterface ist eine KWIC Suche nach Wortformen über den gesamten Korpus oder über Teile des Korpus möglich. Zusätzlich ist die Suche über ein Windowsprogramm (WordCruncher) möglich, das mit einem TITUS-Server in Verbindung steht.

Verfügbarkeit

Online, für den wissenschaftlichen Gebrauch frei.

Projektinformation & Kontakt

Webseite: http://titus.uni-frankfurt.de/

Verantwortlich: Jost Gippert <gippert@em.uni-frankfurt.de>

Literatur

Gippert, Jost (2002): *Der TITUS-Server: Grundlagen eines multilingualen Online-Retrieval-Systems*. In: Gerd Willée / Bernhard Schröder / Hans-Christian Schmitz (Hrsg. / ed.), *Computerlinguistik: Was geht, was kommt? / Computational Linguistics: Achievements and Perspectives. Festschrift für Winfried Lenders* Sprachwissenschaft, Computerlinguistik und Neue Medien, 4, 81-86.

Gippert, Jost (1995): *TITUS. Das Projekt eines indogermanistischen Thesaurus ("TITUS. The project of an Indo-European thesaurus")*. LDV-Forum 12/2, 35-47.

2.5. Codices Electronici Ecclesiae Coloniensis (CEEC)

Textauswahl & Digitalisierungsverfahren

Die Kodices der Erzbischöflichen Diözesan- und Dombibliothek Köln wurden lediglich als Bilddaten eingescannt, sie sind nicht als elektronische Texte verfügbar. Es handelt sich ausschließlich um Handschriften.

Annotation

Es sind Katalogeinträge mit Metadaten (Quellenverweis, etc.) zu den Kodices vorhanden.

Suchmöglichkeiten

Die Suche erfolgt online und über die Katalogeinträge. Das Projekt entwickelt eine Software zur Named Entity Recognition (vor allem zur Erkennung von Personennamen) und ein Werkzeug zur Erstellung paläographischer Dokumentation von Handschriften. Beide sind über die Webseite des Projekts verfügbar.

Verfügbarkeit

Online unter: http://www.ceec.uni-koeln.de

Projektinformation & Kontakt

Webseite: http://www.ceec.uni-koeln.de/

Verantwortlich: Manfred Thaller <manfred.thaller@uni-koeln.de>

Verantwortlich: Heinz Finger <dombibliothek@erzbistum-koeln.de>

2.6. Textkorpus von Thomas Gloning

Textauswahl & Digitalisierungsverfahren

Das Korpus enthält historische und zeitgenössische Texte, von Thomas Gloning digitalisiert. Auswahlkriterien sind nicht angegeben.

Annotation

Die Texte sind nicht positionell oder strukturell annotiert. Sie haben jedoch einen Header, der die genaue Quelle des Textes und Kommentare zur digitalen Fassung enthält.

Suchmöglichkeiten

Es gibt keine Suchmöglichkeiten.

Verfügbarkeit

Das Korpus ist online frei verfügbar.

Projektinformation & Kontakt

Webseite: http://www.staff.uni-marburg.de/~gloning/etexte.htm

Verantwortlich: Thomas Gloning <gloning@mailer.uni.marburg.de>

2.7. Mittelhochdeutsche Begriffsdatenbank (MHDBDB)

Textauswahl & Digitalisierungsverfahren

Die Mittelhochdeutsche Begriffsdatenbank entsteht in einem internationalen Projekt (das seit 1992 läuft) und stellt (im November 2004) ein Korpus aus insgesamt 126 mittelhochdeutschen Texten mit zusammen 5,7 Mio. Wörtern bereit. Es ist möglich nach einzelnen Wörtern, Lemmata und Wortarten zu suchen. Zusätzlich ist eine Konzeptliste/Begriffsliste verfügbar, die es möglich macht, die Texte des Korpus nach bestimmten Konzepten (wie z. B. Werken/Werkzeuge/Allg. Utensilien) zu durchsuchen. Alle digitalisierten Texte entstammen Editionen.

Die MHDBDB ist Grundlage einiger Konzeptwörterbücher.

Die MHDBDB wird laufend erweitert.

Annotation

Die Texte sind morphologisch annotiert, lemmatisiert (und dabei normalisiert) und mit Konzepten annotiert.

Suchmöglichkeiten

Über ein Webinterface ist es möglich in den Texten direkt nach Lemmata oder Wortformen zu suchen. Zusätzlich ist die Suche in der Kategoriendatenbank möglich.

Die Texte haben Headerinformationen und können individuell gruppiert werden.

Verfügbarkeit

Nach einer kostenlosen Anmeldung ist der Onlinezugang und die Suche möglich (auch als anonymer Gast)

Projektinformation & Kontakt

Webseite: http://mhdbdb.sbg.ac.at

Verantwortlich: Klaus M. Schmidt <schmidt@bgnet.bgsu.edu>,

Margarete Springeth < margret.springeth@sbg.ac.at>

3. Historische Korpora anderer Sprachen

3.1. Die Helsinki-Corpus-Familie

Der diachrone Teil des Helsinki Corpus war eines der ersten weit verfügbaren diachronen Korpora und hat die Entwicklung weiterer historischer Korpora mitgeprägt. Viele Methoden der historischen Korpuslinguistik wurden anhand des Helsinki-Korpus entwickelt.

Das ursprüngliche Helsinki Corpus ist nicht positionell annotiert. Inzwischen gibt es einige Projekte, die Texte des Korpus positionell und syntaktisch annotieren. In diesem Abschnitt werden der diachrone Teil des Helsinki Korpus und die daraus entstandenen Korpora vorgestellt.

3.1.1. Diachronic Part of the Helsinki Corpus

Textauswahl & Digitalisierungsverfahren

Das Korpus enthält Texte aus dem Zeitraum 750 – 1700 aus den Sprachstufen Old English, Middle English und Early Modern English. Das Korpus enthält ca. 1,5 Mio Wörter, grob auf die verschiedenen Sprachstufen verteilt. In jeder Sprachstufe wurde auf eine ausgewogene Verteilung zwischen verschiedenen Dialekten und Genres geachtet – um dies zu erreichen, wurden längere Texte nur in Ausschnitten aufgenommen. Die Korpuszusammenstellung wird in Rissanen et al. (1993) und unter http://khnt.hit.uib.no/icame/manuals/HC/INDEX.HTM dargestellt. Der Großteil der Texte ist nach Editionen digitalisiert. Die Texte wurden manuell digitalisiert und mindestens einmal korrekturgelesen.

Die Texte sind in ASCII-Format kodiert. Sonderzeichen und strukturelle Annotation werden durch vordefinierte Zeichenfolgen repräsentiert.

Annotation

Das ursprüngliche Helsinki Corpus enthält keine positionellen linguistischen Annotationen. Es ist aber strukturell und (in geringem Umfang) paläographisch annotiert und mit Headerinformationen versehen. Die Annotationen in diesem Korpus umfassen einige Informationen zur Textausgabe: U. a. Schriftarten, fremdsprachliche Sequenzen, Runen im Original, Emendationen, Anmerkungen des Editors und des Korpuserstellers, Überschriften, Zeilen und Absätze.

Metainformationen zum Text (Headerinformationen) sind nach den COCOA Konventionen annotiert. Die Annotationen umfassen 25 Attribute, darunter: Autor, Geschlecht, Alter und sozialen Rang des Autors, Datum des Originals und des Manuskripts, eventuelle Beziehung zu fremdsprachlichen Originalen und prototypische Textkategorie.

Dazu siehe auch: http://khnt.hit.uib.no/icame/manuals/HC/INDEX.HTM#con334

Die Annotation erfolgte manuell.

Suchmöglichkeiten

Das Korpus kann mit Hilfe des Oxford Concordance Program und des (kostenfrei erhältlichen) Programms WordCruncher durchsucht werden.

Verfügbarkeit

Das Korpus ist zusammen mit der ICAME Corpus Collection auf CD-ROM erhältlich, siehe http://www.hit.uib.no/icame/cd/

Der Erwerb und Gebrauch der ICAME Corpus Collection ist nur zu Forschungszwecken gestattet.

Projektinformation & Kontakt

The HIT Centre/Humanities Information Technologies Research Programme
Allégaten 27, N-5007 Bergen, Norway
E-mail: icame@hit.uib.no
URL: http://www.hit.uib.no/
Tel: +47 55 582954/55/56
Fax: +47 55 589470

Literatur

Kytö, Merja (comp.) (1996). *Manual to the diachronic part of the Helsinki Corpus of English Texts. Coding conventions and lists of source texts.* 3rd edition. Department of English, University of Helsinki. Im Internet unter http://khnt.hit.uib.no/icame/manuals/HC/INDEX.HTM.

Rissanen, Matti; Kytö, Merja & Pallander, Minna (1993) *Early English in the Computer Age: Explorations through the Helsinki Corpus* Mouton de Gruyter, Berlin

Eine Reihe von Publikationen zur Auswertungen des Helsinki Corpus findet sich unter http://www.eng.helsinki.fi/varieng/team1/1_1_2_publications.htm

Spezielles

Inzwischen gibt es linguistisch annotierte Teile des Helsinki Corpus:

- das York-Helsinki Parsed Corpus of Old English Poetry (York Poetry Corpus)
- das York-Toronto-Helsinki Parsed Corpus of Old English Prose (YCOE)
- das Penn-Helsinki Parsed Corpus of Middle English, second edition (PPCME2)

Alle Teilkorpora sind nach den gleichen Prinzipien morphologisch und syntaktisch annotiert. Die Annotation ist an die Penn Treebank[6] angelehnt. Alle drei Korpora können mit dem gleichen Suchprogramm CorpusSearch durchsucht werden. Im folgenden werden die drei Korpora einzeln beschrieben.

[6] Dazu siehe http://www.cis.upenn.edu/~treebank/home.html

3.1.2. Brooklyn-Geneva-Amsterdam-Helsinki Parsed Corpus of Old English (Brooklyn Corpus)

Textauswahl & Digitalisierungsverfahren

Das Brooklyn Corpus ist ein Vorläufer des York-Toronto-Helsinki Parsed Corpus of Old English Prose (YCOE). Da das YCOE umfangreicher und tiefer annotiert ist, beschreiben wir das Brooklyn Corpus hier nur kurz. Die Texte, die in diesem Korpus enthalten sind, bilden eine Auswahl des altenglischen Teils des Helsinki Corpus. Insgesamt umfasst das Korpus 106210 Wörter.

Annotation

Das Korpus ist syntaktisch und morphologisch annotiert.

Suchmöglichkeiten

Es sind vier Formate von jedem Text verfügbar, für verschiedene Suchmöglichkeiten: Einfache Textsuche, Suche mit eigenen PERL-Skripten oder CorpusSearch (PPCME2)

Verfügbarkeit

Die Texte sind nach einer Anmeldung für den akademischen Gebrauch frei benutzbar, die Rechte für die Texte liegen beim Helsinki Korpus. Die Rechte für die Annotation liegen bei Susan Pintzuk und Eric Haeberli.

Projektinformation & Kontakt

Webseite: http://www-users.york.ac.uk/~sp20/corpus.html

Verantwortlich: Susan Pintzuk <sp@york.ac.uk>

Literatur

Auf den Seiten des Korpus (http://www-users.york.ac.uk/~sp20/corpus.html) gibt es das Handbuch dazu in mehreren Formaten.

3.1.3. The York-Toronto-Helsinki Parsed Corpus of Old English Prose (YCOE)

Textauswahl & Digitalisierungsverfahren

Das YCOE ist ein syntaktisch annotiertes Korpus altenglischer Prosatexte. Es umfasst ca. 1,5 Millionen Wörter. Die aufgenommenen Texte stammen zum Teil aus dem Helsinki Corpus.

Annotation

Das Korpus ist strukturell und positionell annotiert (Flexionsmorphologie und Syntax) und mit Headerinformation versehen. Das Annotationsformat ist am Format der Penn Treebank orientiert.

Suchmöglichkeiten

Die Suche nach syntaktischen Strukturen wird durch das Programm CorpusSearch ermöglicht.

Verfügbarkeit

Das Korpus ist über das Oxford Text Archive für nicht kommerzielle Zwecke kostenlos erhältlich.

Projektinformation & Kontakt

Webseite: http://www-users.york.ac.uk/~lang22/YcoeHome1.htm

Verantwortlich: Ann Taylor <at9@york.ac.uk>

3.1.4. The York-Helsinki Parsed Corpus of Old English Poetry (York Poetry Corpus)

Textauswahl & Digitalisierungsverfahren

Das York Poetry Corpus ist ein syntaktisch annotiertes Korpus bestehend aus einer Auswahl von poetischen Texten aus dem altenglischen Abschnitt des Helsinki Corpus. Das Korpus umfasst insgesamt 71490 Wörter.

Annotation

Das Korpus ist strukturell und positionell annotiert (Flexionsmorphologie und Syntax) und mit Headerinformation versehen. Das Annotationsformat ist am Format der Penn Treebank orientiert.

Suchmöglichkeiten

Die Suche nach syntaktischen Strukturen ist mit CorpusSearch möglich.

Verfügbarkeit

Das York Poetry Corpus ist für Lehr – und Forschungszwecke frei zugänglich. Der Zugang erfolgt entweder direkt über Susan Pintzuk oder über das Oxford Text Archive (OTA).

Projektinformation & Kontakt

Webseite: http://www-users.york.ac.uk/~lang18/pcorpus.html

Verantwortlich: Susan Pintzuk <sp@york.ac.uk>

Literatur

Es gibt ein Handbuch, das online zugänglich ist (http://www-users.york.ac.uk/~lang18/Documentation/Contents-oe.htm)

3.1.5. Penn-Helsinki Parsed Corpus of Middle English

Textauswahl & Digitalisierungsverfahren

Das Korpus enthält Textsamples aus dem Middle English Abschnitt des Helsinki Corpus und weitere mittelenglische Texte (dazu gibt es keine Angaben zur Digitalisierung). Es hat insgesamt einen Umfang von 1.3 Millionen Wortformen aus 55 Texten.

Annotation

Das Korpus ist strukturell und positionell annotiert (Flexionsmorphologie und Syntax) und mit Headerinformation versehen. Das Annotationsformat ist am Format der Penn Treebank orientiert.

Suchmöglichkeiten

Die Suche nach syntaktischen Strukturen ist mit CorpusSearch möglich.

Verfügbarkeit

Eine CD-ROM mit PPCME2 und CorpusSearch kann bestellt werden (50$), Bestellformulare und Anschrift unter: http://www.ling.upenn.edu/mideng/ppcme2dir/order-forms.html

Projektinformation & Kontakt

Webseite: http://www.ling.upenn.edu/mideng/

Kontakt: <ppcme2@babel.ling.upenn.edu>

Fragen zu CorpusSearch: <corpus-search@babel.ling.upenn.edu>

3.2. Weitere Historische Korpora des Englischen

3.2.1. Lancaster Newsbook Corpus

Textauswahl & Digitalisierungsverfahren

Newsbooks sind ein Vorläufer der modernen Zeitungen aus dem 17. Jh. Solche Texte bestehen oft aus Textbausteinen, die aus anderen Newsbook-Texten übernommen wurden. Um verschiedene Newsbooks quantitativ mit einander vergleichen zu können wurde das Lancaster Newsbook Corpus aufgebaut.

Das Korpus besteht aus englischen Newsbook-Texten aus den Thomason Tracts, die von Dezember 1653 bis Mai 1654 veröffentlicht wurden. Der Gesamtumfang beträgt 800.000 Wörter.

Alle Texte wurden von Hand digitalisiert. Als Vorlage dienten hierbei Mikrofilmkopien der Originale.

Um den Vergleich der Texte zu ermöglichen wurde die Orthographie der Texte normalisiert.

Annotation

Die Texte sind nach dem TEI SGML Standard annotiert. Es werden leider keine Angaben gemacht welche Informationen zu den einzelnen Texten annotiert sind.

Suchmöglichkeiten

Keine Angaben

Verfügbarkeit

Der erste Release des Korpus ist auf CD-ROM erhältlich.

Projektinformation & Kontakt

Webseite: http://www.ling.lancs.ac.uk/newsbooks/default.htm

Verantwortlich: Tony McEnery <a.mcenery@lancaster.ac.uk>

Vertrieb: Dawn Archer <d.archer@lancaster.ac.uk>

Literatur

Paul Clough, Robert Gaizauskas, S. L. Piao (2002): *Building and annotating a corpus for the study of journalistic text reuse*. In Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC-02), pp.1678-1691 (Vol V), 29-31st May 2002, Los Palmas de Gran Canaria, Spain.

3.2.2. Corpus of Late 18c Prose

Textauswahl & Digitalisierungsverfahren

Das Korpus enthält einige Briefe, datiert 1761-1789, aus der 'Leghs of Lyme' Kollektion. Es sind Briefe von verschiedenen Verfassern an Richard Orford, einen Gutsverwalter. Die Verfasser der Briefe gehören verschiedenen sozialen Klassen an.

Keine Angaben zu Digitalisierungsverfahren oder dem Umfang des Korpus.

Annotation

Die Texte sind im COCOA Stil des Helsinki Corpus annotiert. Es sind Informationen über Autor, Datum und Seitenumbrüche verfügbar.

Suchmöglichkeiten

Keine Angaben.

Verfügbarkeit

Die Texte sind für Lehr- und Forschungszwecke kostenlos online zugänglich, nachdem eine Lizenzvereinbarung unterschrieben wurde. Das Copyright für die Texte liegt bei der John Rylands University Library of Manchester, für die Annotationen liegt es bei David Denison und Linda van Bergen.

Das Korpus ist online über das Oxford Text Archive oder über die Webseite des Projekts verfügbar.

Die Zugangsberechtigung kann durch Ausfüllen eines Formulars erworben werden.

Projektinformation & Kontakt

Webseite: http://www.art.man.ac.uk/english/staff/dd/late18c.htm

Zugangsformular: http://www.art.man.ac.uk/english/staff/dd/late18c_access_request.txt

Verantwortlich: David Denison <d.denison@man.ac.uk>

3.2.3. Corpus of Early English Correspondence (CEEC), CEEC Extensions, CEEC Supplement

Textauswahl & Digitalisierungsverfahren

Das CEEC wurde von einer Forschergruppe der Universität Helsinki, bestehend aus Soziolinguisten und historischen Linguisten zwischen 1993 und 1998 erstellt. Zu diesem Zeitpunkt bestand das Korpus aus circa 2,7 Millionen Wörtern, die aus Privatbriefen aus dem Zeitraum 1471 bis 1681 entnommen wurden. Insgesamt sind 6000 Briefe enthalten die von 800 verschiedenen Autoren verfasst wurden. Zusätzlich zu den Texten wurden verschiedene soziolinguistische Parameter zu den Autoren aufgenommen. Das CEEC von 1998 ist ausgewogen bezüglich dieser Parameter. Deshalb wurde beschlossen dieses Korpus nicht zu erweitern.

Stattdessen wurden 2000 zwei Nachfolgeprojekte ins Leben gerufen, deren Aufgabe es ist, jeweils ein Erweiterungskorpus zu erstellen. Einerseits die CEEC Extension mit Privatbriefen von 1681 bis 1800. Andererseits das CEEC Supplement, das Privatbriefe enthalten soll, die nicht den statistischen Kriterien des übrigen Korpus entsprechen.

Die Briefe im 1998er CEEC und in der CEEC Extension wurden nach den gleichen Kriterien aufgenommen. Wichtigstes Kriterium ist die Authentizität der Briefe. Wenn möglich, wurden die Texte aus vorhandenen Editionen digitalisiert. Im Idealfall wurden solche Editionen gewählt, die genaue Angaben über den Verfasser eines Briefes enthielten.

Ein zweites wichtiges Kriterium ist die Ausgewogenheit der Zusammenstellung der Texte. So wurde versucht Briefe von männlichen und weiblichen Verfassern aus allen sozialen Schichten aufzunehmen. Von jedem Verfasser wurden wenn möglich, 10 mittellange Briefe aufgenommen.

Die aufgenommenen Texte wurden nicht orthographisch normalisiert. Diese Entscheidung wurde getroffen, um die Untersuchung von orthographischen Unterschieden zu ermöglichen.

Annotation

Die Kodierung und Annotation des CEEC hält sich an die COCOA Konvention des Helsinki Corpus. Es sind vor allem Informationen zu den Verfassern annotiert, außerdem Informationen zur Textstruktur. Linguistische Annotationen sind nicht vorhanden. Die Annotation der Texte wurde von Hand durchgeführt.

Suchmöglichkeiten

Die Texte können mit dem Oxford Concordance Program analysiert werden.

Verfügbarkeit

Das Korpus ist zusammen mit der ICAME Corpus Collection auf CD-ROM erhältlich, siehe http://www.hit.uib.no/icame/cd/

Der Erwerb und Gebrauch der ICAME Corpus Collection ist nur zu Forschungszwecken gestattet.

Projektinformation & Kontakt

Webseite: http://www.eng.helsinki.fi/varieng/team2/1_2_4_projects.htm

Verantwortlich: Terttu Nevalainen < terttu.nevalainen@helsinki.fi>

Literatur

Laitinen, Mikko (2002). *Extending the Corpus of Early English Correspondence to the 18th Century*: Online unter http://www.eng.helsinki.fi/hes/Corpora/extending_the_corpus.htm.

Keränen, Jukka (1998). *The Corpus of Early English Correspondence: Progress report* in *Explorations in Corpus Linguistics* ed. by Antoinette Renouf. Amsterdam: Rodopi. 29-37.

Nevalainen, Terttu & Helena Raumolin-Brunberg, (Hrsgb.) (1996). *Sociolinguistics and language history. Studies based on the Corpus of Early English Correspondence*. (Language and computers 15.) Amsterdam & Atlanta: Rodopi

3.2.4. Lampeter Corpus of Early Modern English

Textauswahl & Digitalisierungsverfahren

Das Lampeter Corpus enthält Traktate und Pamphlete, die zwischen 1640 und 1740 veröffentlicht (gedruckt) wurden und in der Founders' Library der University of Wales, Lampeter aufbewahrt werden. Für jede Dekade dieses Zeitraums wurden aus dem 6 Domänen Religion, Politik, Wirtschaft & Handel, Wissenschaft, Recht und Verschiedenes jeweils zwei Texte ausgewählt. Von einem Autor wurde jeweils nur ein Text aufgenommen. Texte von sehr bekannten Autoren wurden ausgelassen. Insgesamt enthält das Korpus 120 Texte mit ca. 1,1 Millionen laufenden Wortformen.

Es sind jeweils komplette Texte in der Erstausgabe mit Widmungen, Vorwörtern, etc. Spätere Ausgaben eines Textes wurden nur verwendet, wenn keine Änderungen am eigentlichen Text gemacht wurden.

Annotation

Die Texte liegen im TEI SGML Format vor. Es sind Informationen über Autor, Drucker/Herausgeber, Datum des Drucks, Publikationsformat, Textcharakteristik und bibliographische Bezüge annotiert. Zu einigen Texten gibt es Faksimiles.

Suchmöglichkeiten

Die Webseite des Projekts stellt die Texte aus dem Korpus in TEI SGML kodiert und als Präsentationsversion zur Verfügung. Suchmöglichkeiten scheinen nicht zu bestehen.

Verfügbarkeit

Das Lampeter Corpus kann direkt über die Webseite des Projekts erworben werden. Zusätzlich kann man das Korpus auch über das Oxford Text Archive oder ICAME erwerben.

Projektinformation & Kontakt

Webseite:
http://www.tu-chemnitz.de/phil/english/chairs/linguist/real/independent/lampeter/lamphome.htm

Verantwortlich: Rainer Siemund (REAL Centre TU-Chemnitz):
<real.centre@phil.tu-chemnitz.de>

Literatur:

Manual (Claudia Claridge):
http://www.tu-chemnitz.de/phil/english/chairs/linguist/real/independent/lampeter/manual/pages/manual.html

Siemund, Rainer & Claudia Claridge. 1997. *The Lampeter Corpus of Early Modern English tracts*. in: *ICAME Journal* 21: 61-70.

3.2.5. Corpus of Middle English Prose and Verse

Textauswahl & Digitalisierungsverfahren

Das Korpus enthält Lyrik und Prosatexte aus der mittelenglischen Periode, die für die Erstellung des Middle English Dictionary (MED) der Humanities Text Initiative (HTI) der Universität Michigan verwendet wurden. Die Texte sind aus Editionen übernommen, die vom Oxford Text Archive bereitgestellt wurden. Eine Aufstellung der Texte befindet sich unter: http://www.hti.umich.edu/c/cme/bibl.html.

Annotation

Die Texte in diesem Korpus sind im TEI SGML Format annotiert (TEI Lite DTD).

Es sind einige Metainformationen zu Text & Edition annotiert. Unter anderem Titel, Editor, Vertreiber und Verleger zu jeder Ausgabe des Texts. Anmerkungen des Editors zu strittigen Stellen sowie zu jedem Text.

Suchmöglichkeiten

Zu diesem Korpus existiert ein Webinterface das verschiedene Suchfunktionen bereitstellt.

Simple Search ermöglicht die Suche nach einzelnen Wörtern und Phrasen im gesamten Korpus. Boolean Search ermöglicht die Suche nach zwei Wörtern die zusammen in einer Zeile einem Absatz oder einem Werk vorkommen. Proximity Search stellt eine Suchfunktion nach bis zu 3 Begriffen die nah beieinander vorkommen dar.

Verfügbarkeit

Das Webinterface ist frei zugänglich. Zur Verfügbarkeit des Korpus an sich werden keine Angaben gemacht.

Projektinformation & Kontakt

Dieses Korpus ist ein Projekt der HTI.

Kontaktadresse: <hti-info@umich.edu>

Webseite: http://www.hti.umich.edu/c/cme/.

3.2.6. Innsbrucker Computer-Archive of Machine-Readable English Texts

Textauswahl & Digitalisierungsverfahren

Das Korpus besteht aus zwei Teilen. Der erste Teil ist das Prose Corpus of ICAMET: Er enthält 129 Prosatexte in Middle English aus dem Zeitraum 1100-1500, die von vorhandenen Editionen digitalisiert wurden.

Der zweite Teil ist das Letter Corpus of ICAMET: Er enthält 254 komplette Briefe, verfasst zwischen 1386 und 1688.

Annotation

Keine Angaben

Suchmöglichkeiten

Keine Angaben

Verfügbarkeit

Dieses Korpus ist zusammen mit der ICAME Corpus Collection auf CD-ROM erhältlich, siehe http://www.hit.uib.no/icame/cd/

Der Erwerb und Gebrauch der ICAME Corpus Collection ist nur zu Forschungszwecken gestattet.

Das Benutzerhandbuch zu diesem Korpus befindet sich nicht auf der ICAME CD-ROM, sondern muss separat über das English Department der Universität Innsbruck erworben werden.

Projektinformation & Kontakt

Webseite: http://anglistik1.uibk.ac.at/ahp/projects/icamet/prose_corpus/index.html

Verantwortlich: <Manfred.Markus@uibk.ac.at>

3.2.7. Renaissance Electronic Texts (RET)

Textauswahl & Digitalisierungsverfahren

Das RET Korpus ist eine Sammlung von Texten aus dem 16. und 17. Jahrhundert die in Early Modern English verfasst sind. Das Korpus besteht aus einer Sammlung von drei Editionen die von RET selbst erarbeitet wurden, zwei Einzeltexten, die ebenfalls von diesem Projekt editiert wurden, und einer Datenbank von zweisprachigen Wörterbüchern und Lexika.

Die Texte wurden zum Zweck der literaturwissenschaftlichen Forschung ausgewählt

Annotation

Die Texte sind in einem SGML Format enkodiert, dass in den RET Encoding Guidelines beschrieben ist. Insbesondere wird hier die Entscheidung gegen die TEI Guidelines erläutert.

Alle Texte wurden mit Programm Pat der Open Text Corporation indiziert.

Suchmöglichkeiten

Suche nach Wortformen ist über ein Webinterface möglich. Dabei kann ein Wort, das sich im Kontext des gesuchten Wortes finden soll, als zusätzliches Kriterium angegeben werden.

Verfügbarkeit

Das Korpus ist für Forschungszwecke nach Registrierung online verfügbar.

Projektinformation & Kontakt

Webseite: http://www.library.utoronto.ca/utel/ret/ret.html

Verantwortlich: Ian Lancashire <ian@chass.utoronto.ca>

Literatur

Richtlinien zur Textenkodierung der Renaissance Electronic Texts:

Ian Lancashire 1994. *RET Encoding Guidelines.*
http://www.library.utoronto.ca/utel/ret/ret.html

3.2.8. A Glossarial DataBase of Middle English

Textauswahl & Digitalisierungsverfahren

Das Korpus gehört zu einer größeren Glossardatenbank für Middle English.

Keine Angaben zu Digitalisierungsverfahren oder dem Umfang des Korpus.

Annotation

Die Texte in diesem Korpus wurden von Hand nach dem TEI-Standard für analytische Annotationen annotiert. Die Annotation umfasst Lemma und Part-of-Speech Informationen. Das verwendete Tagset ist online dokumentiert.

Suchmöglichkeiten

Über ein Webinterface ist eine Volltextsuche oder eine Suche nach Lemma oder Part-of-Speech möglich. Dabei wird jeweils nur ein Satz ausgegeben (der Kontext kann nicht erweitert werden).

Verfügbarkeit

Das Suchinterface ist frei verfügbar. Genauere Angaben zu den Zugangsbedingungen für das Korpus selbst werden nicht gemacht.

Projektinformation & Kontakt

Webseite: http://www.hti.umich.edu/g/gloss/

Verantwortlich: Larry Benson <ldb@wjh12.harvard.edu>

3.2.9. The Canterbury Tales Project

Textauswahl & Digitalisierungsverfahren

Das Canterbury Tales Project wurde mit der Zielsetzung ins Leben gerufen, die Textgeschichte der Canterbury Tales mit computerbasierten Methoden zu rekonstruieren. Die Canterbury Tales sind ein Text von Geoffrey Chaucer, den er vor seinem Tod 1400 nicht mehr vollenden konnte. Es existieren jedoch 88 verschieden abgewandelte Fassungen des Textes. In den 1930ern wurden alle diese Fassungen von John Manly und Edith Rickert gesammelt. Mit Hilfe von Kollationskarten wurden sie wortweise verglichen, um eine Aussage über die Textgeschichte zu machen. Die Ergebnisse dieser Arbeit sind 1941 in einer Edition veröffentlicht worden. Diese Edition fand jedoch nie richtige Akzeptanz, da die Methoden von Manly und Rickert kritisierbar waren.

Das Canterbury Tales Project hat alle Fassungen des Textes manuell digitalisiert.

Annotation

Die digitalisierten Texte wurden mit Hilfe der Software Collate normalisiert und kollationiert. Kollationierung meint eine Wort-zu-Wort-Alignierung aller Textfassungen.

Suchmöglichkeiten

Die Textgeschichte wird mit Hilfe sog. phylogenetischer Algorithmen untersucht. Das sind Algorithmen aus der evolutionären Biologie, die dort zur Berechnung von Stammbäumen biologischer Arten verwendet werden. Das Canterbury Tales Project verwendet zwei verschiedene Programme die zu diesem Zweck entwickelt wurden. PAUP (Phylogenetic Analysis using Parsimony) und SPLITSTREE.

Die Kollationierten Textfassungen sind in einer Datenbank erfasst, auf die mit einer eigenen Anfragesprache zugegriffen werden kann.

Verfügbarkeit

Das Projekt gibt Texteditionen der Canterbury Tales heraus, die auf den Erkenntnissen der Forschungsarbeit des Projekts beruhen. Diese Editionen sind auf CD-ROM über Cambridge University Press erhältlich. Die CD-ROM-Editionen erhalten jeweils alle zugrundeliegenden Textversionen aus dem Korpus.

Projektinformation & Kontakt

Webseite: http://www.cta.dmu.ac.uk/projects/ctp/index.html

Verantwortlich: Peter Robinson <peter.robinson@dmu.ac.uk>

Literatur

Norman Blake, Peter Robinson (Hrsg.).(1993-1997): *The Canterbury Tales Project: Occasional Papers*. (Vol I – II). Office for Humanities Communication, Centre for Computing in the Humanities, King's College London.

3.3. Korpora des Lateinischen

3.3.1. Corpus Scriptorum Latinorum (CSL)

Textauswahl & Digitalisierungsverfahren

Das Corpus Scriptorum Latinorum ist ein Index aller online verfügbaren lateinischen Texte und zugehörigen Übersetzungen. Eine Aufstellung aller enthaltenen Texte befindet sich unter http://www.forumromanum.org/literature/table.html.

Annotation

Die Texte sind mit den Metainformationen Titel, Autor und Edition versehen.

Suchmöglichkeiten

Es können ganze Texte über die Aufstellung (s. o.) gesucht werden. Zusätzlich gibt es eine nach Autoren geordnete Aufstellung; nach Titel, Genre und Datum geordnete Aufstellungen sind in Arbeit.

Zusätzlich kann das Vorkommen von Wörtern in den Texten mit Hilfe des webbasierten Tools Picosearch gesucht werden. Die Vorkommen können jedoch nicht im Kontext angezeigt werden (sondern nur die Texte in denen sie sich befinden, wie bei Google).

Verfügbarkeit

Online unter http://www.forumromanum.org/literature/index.html.

Die rechtliche Seite der Verfügbarkeit ist für jeden Text einzeln geregelt, je nach den Vorstellungen des Editors. Einige enthaltene Texte sind jedoch frei verfügbar.

Projektinformation & Kontakt

Webseite: http://www.forumromanum.org/literature/index.html

Verantwortlich: David Camden <camden@fas.harvard.edu>

3.3.2. La Banque de Texte de LASLA

Textauswahl & Digitalisierungsverfahren

Das Korpus enthält folgende lateinische Texte von Cicero:

Pro A. Caecina or., *Divination in C.*, *Pro M. Fonteio or.*, *De Imperio Cn. Pomp*, *Pro P. Quinctio*, *Pro S. Roscio Amer.*, *Pro Q. Roscio Como.*, *Pro M. Tullio or.*, *In C. Verrem I*, *In c. Verrem II*, *De Amicitia*, *De Officiis*, *De Senectute*.

Die Texte sind aus nicht näher bestimmten Editionen übernommen.

Annotation

Das Korpus ist mit Lemmainformationen annotiert. Die Lemmatisierung wurde mit Hilfe des Wörterbuchs *Lexicon totius latinitatis* (Forcellini, Coradini, Padoue 1864) durchgeführt. Dazu wurden die Lemmanamen des Wörterbuchs normalisiert und einige Lemmata hinzugefügt. Ambige Lemmata wurden mit zusätzlichen Informationen versehen, entweder der Wortart oder einer Beschreibung des Sinns des Lemmas.

Die im Text enthaltenen Wortformen wurden normalisiert und mit einigen morphologischen Informationen annotiert: Wortart, Flexionsklasse, Kasus, Numerus, Steigerungsform, Aktiv/Passiv, Modus, Tempus, Person.

Zusätzlich ist zu jedem Verb annotiert, ob es ein Hauptverb oder untergeordnetes Verb ist.

Suchmöglichkeiten

Die Suche ist über ein Webinterface möglich.

Es kann nach allen annotierten Informationen zu einer einzelnen Wortform gesucht werden oder zu mehrere Wortformen im Kontext.

Verfügbarkeit

Das Korpus steht nach Einschreibung für eine Testphase gratis zur Verfügung. Weitere Modalitäten sind nicht ersichtlich.

Projektinformation & Kontakt

Le Laboratoire d'Analyse Statistique des Langues Anciennes
http://www.ulg.ac.be/cipl/bdlasla/

Verantwortlich: Joseph Denooz <Joseph.Denooz@ulg.ac.be>

3.3.3. Augustana

Textauswahl & Digitalisierungsverfahren

Das Korpus enthält eine Zusammenstellung literarischer Texte in verschiedenen europäischen Sprachen die aus verschiedenen Editionen übernommen wurden. Enthalten sind: Lateinische Texte vom 7. Jh. v. Chr. bis zum 19. Jh., griechische Texte vom 8. Jh. v. Chr. bis zum 15. Jh., deutsche Texte vom 8. Jh. bis zum 20. Jh., englische Texte vom 9. bis zum 20. Jh., französische Texte vom 9. - 20. Jh., italienische Texte vom 9. - 20. Jh., spanische Texte vom 11. - 20. Jh. und polnische Texte vom 12. - 20. Jh.

Es werden keine Angaben zum Digitalisierungsverfahren gemacht; alle Texte sind im HTML Format aufbereitet

Annotation

Das Korpus ist nicht linguistisch annotiert. Metainformation zu Autor, Edition und Originalform des Textes sind verfügbar.

Suchmöglichkeiten

Es gibt keine Suchmöglichkeiten. Das Korpus besteht aus HTML Seiten. Die Navigation wird dadurch erschwert, dass sie in Latein gehalten ist.

Verfügbarkeit

Das Korpus ist online frei verfügbar.

Projektinformation & Kontakt

Webseite: http://www.fh-augsburg.de/~harscg/augustana.html

Verantwortlich: Ulrich Harsch <harsch@rz.fh-augsburg.de>

3.4. Historische Korpora des Spanischen und des Portugiesischen

3.4.1. Corpus del Espanol

Textauswahl & Digitalisierungsverfahren

Das Corpus del Espanol ist ein diachrones Korpus des Spanischen. Es enthält 100 Millionen Wörter, 20 Millionen aus dem Zeitraum 1100-1400, 40 Millionen aus dem Zeitraum 1400-1700 und 40 Millionen aus dem Zeitraum 1700-1900.

Die Texte aus dem 19. Jahrhundert sind gleichmäßig aus Literatur, gesprochenen Texten und Zeitungen/Enzyklopädien zusammengesetzt.

Annotation

Das Korpus ist mit Part-of-Speech und Lemmainformationen annotiert. Zusätzlich sind Frequenzinformationen zu allen eindeutigen 1-, 2- und 3-Wort Sequenzen enthalten. Es können beliebige andere Informationen hinzugefügt werden. Dies wird durch eine relationale Datenbank ermöglicht.

Es gibt keine Angaben dazu, wie die Annotation durchgeführt wurde.

Suchmöglichkeiten

Über ein Webinterface kann nach Form, Lemma, grammatischer Kategorie oder Frequenz eines Tokens im gesamten Korpus gesucht werden. Die Anzeige des Ergebnisses erfolgt gruppiert nach Zeitperioden. Die Vorkommen können dann im Kontext angezeigt werden.

Verfügbarkeit

Das Webinterface ist frei zugänglich. Ansonsten gibt es keine Distribution des Korpus.

Projektinformation & Kontakt

Projektwebseite: http://www.corpusdelespanol.org

Verantwortlich: Prof. Mark Davies <mark_davies@byu.edu>

3.4.2. Tycho Brahe Parsed Corpus of Historical Portuguese

Textauswahl & Digitalisierungsverfahren

Das Korpus beinhaltet Texte von portugiesischen Autoren der Zeit 1550 bis 1850. Es handelt sich um 41 Texte mit ca. 1,8 Millionen laufenden Wortformen. Zu den Auswahlkriterien werden keine Angaben gemacht. Alle Texte sind orthographisch normalisiert.

Annotation

Die Texte sind morphologisch und syntaktisch annotiert, nach einem Schema von Helena Britto und Charlotte Galves, das sich stark an dem Schema des Helsinki Corpus orientiert.

Suchmöglichkeiten

Nach Registrierung ist der Zugriff auf die Texte über ein Webinterface möglich. Weitere Suchmöglichkeiten sind nicht ersichtlich.

Verfügbarkeit

Das Korpus ist online verfügbar nach einer kostenlosen Registrierung.

Projektinformation & Kontakt

Webseite: http://www.ime.usp.br/~tycho/corpus/

Verantwortlich: Charlotte Galves <galvesc@ime.usp.br>

Literatur

Handbuch zur Annotation: http://www.ime.usp.br/~tycho/corpus/manual/index.html

3.5. Historische Korpora des Französischen

3.5.1. Textes de Français Ancien (TFA)

Textauswahl & Digitalisierung

Das Korpus enthält Texte in Alt- und Mittelfranzösisch aus dem 13. – 15. Jahrhundert. Es umfasst insgesamt ca. 3 Millionen laufende Wortformen verteilt auf ca. 120 Texte.

Die Texte wurden möglichst eng an vorliegenden Editionen oder Handschriften transkribiert. Emendationen wurden jedoch in eckigen Klammern zum Text hinzugefügt.

Annotation

Es sind Informationen über die Seitenzahlen der zugrunde liegenden Handschriften oder Editionen enthalten.

Suchmöglichkeiten

Es gibt ein Webinterface zur Volltextsuche in den Texten.

Verfügbarkeit

Das Korpus ist für wissenschaftliche Zwecke über ATILF- *Analyse et Traitement Informatique de la Langue Française* verfügbar (siehe http://www.atilf.fr)

Projektinformationen & Kontakt

Webseite: http://www.uottawa.ca/academic/arts/lfa/

Verantwortlich: Pierre Kunstmann < kunstman@uottawa.ca>

3.5.2. Frantext

Textauswahl & Digitalisierung

Das Frantext-Korpus enthält ca. 3500 französische Texte aus dem 16. - 20. Jh. Es sind 80 % literarische Texte und 20 % technische Texte.

Zum Digitalisierungsverfahren werden keine Angaben gemacht.

Annotation

Zu allen Texten im Korpus sind bibliographische Informationen verfügbar. Ein Teil der Texte des Korpus sind mit Annotationen zur Wortart und zu grammatikalischen Rollen versehen.

Suche

Es existiert eine bibliographische Datenbank zu diesem Korpus, in der nach Titeln oder Autoren gesucht werden kann.

Ein Webinterface bietet sehr umfangreiche Arbeitsmöglichkeiten für linguistische Fragestellungen. Der Benutzer muss sich zunächst ein Arbeitskorpus definieren und kann verschiedene Funktionen auf diesem Arbeitskorpus ausführen:

- Suche von einzelnen Worten und Wortfolgen und deren Anzeige im Kontext.
- Erstellen von Wortlisten, die zur Suche, zur Frequenzanalyse und zur Konkordanzerstellung verwendet werden können. Wortlisten können manuell, über Flexionsklassen oder über Muster von Zeichenfolgen erstellt werden.
- Frequenzanalyse: Absolute und relative Frequenzen von graphischen Wortformen
- Erstellen von kontextfreien Grammatiken, die als Suchanfrage verwendet werden können

Verfügbarkeit

Die Rechte an den Texten selbst werden entweder durch die Öffentlichkeit gehalten (Public Domain), oder es gibt keine Rechtehalter.

Das Korpus ist Abonnementen im Internet zugänglich. Ein Abonnement kostet 35 € pro Jahr für Einzelpersonen und 310 € pro Jahr für wissenschaftliche Einrichtungen. Ein Abonnement für Einrichtungen erlaubt den gleichzeitigen Zugriff von bis zu 50 Personen.

Projektinformation & Kontakt

Base textuelle FRANTEXT
CNRS - ATILF
B.P. 30687
54063 NANCY Cedex
Tel. +33 (0) 3 83 96 21 76
Fax +33 (0) 3 83 97 24 56
Internet: http://www.atilf.fr/

3.6. Historische Korpora von slawischen Sprachen

3.6.1. Annotated Corpora of Text (ACT)

Textauswahl & Digitalisierung

Das Korpus enthält 18 altkirchenslawische Texte aus dem Zeitraum 1230 bis 1450 und umfasst ca. 700 000 laufende Wortformen. Die Texte wurden größtenteils aus einem Vorläuferprojekt übernommen, über das Digitalisierungsverfahren werden keine Angaben gemacht.

Das Korpus ist in einer relationalen Datenbank gespeichert. Die Texte werden aus einem eigens entwickelten XML-Format importiert und können wieder in dieses Format exportiert werden. Bei den Altkirchenslawischen Texten handelt es sich um Handschriften und damit ist die digitale Repräsentation der Texte nicht einfach. Die Repräsentation der Texte in der Datenbank und im Import-/ Exportformat behandelt Sonderzeichen, fehlende Trennzeichen zwischen den Wortformen und Abkürzungen. Dazu werden die Originalform und eine normalisierte Form jeder Wortform repräsentiert.

Annotation

Das Korpus enthält positionelle Attribute für die Annotationsebenen (normalisierte) Wortform, Lemma, Wortart und Morphologie. Strukturell werden Seite und Zeile annotiert.

Suchmöglichkeiten

Die Suche im Korpus wird über ein Webinterface ermöglicht. Über einen Katalog können die enthaltenen Texte entweder direkt betrachtet werden oder nach bestimmten Werten auf den einzelnen Annotationsebenen durchsucht werden. Bei der Suche können bestimmte typische Wortformalternativen eingeschlossen werden.

Alternativ zu diesem Webinterface gibt es auch eine Java-Software, die die Abfrage derselben Informationen ermöglicht. Es können sowohl lokal als auch entfernt installierte Korpora abgefragt werden.

Verfügbarkeit

Das Korpus ist direkt im Internet verfügbar (siehe

http://prometheus.ms.mff.cuni.cz/act/www/index.php). Das gesamte Korpus, die zugehörige Java-Software und das Webinterface sind auch zum Download verfügbar. Die Software steht unter einer Open-Source Lizenz, so dass sie auch von anderen Parteien benutzt und angepasst werden kann.

Projektinformationen & Kontakt

Verantwortlich:
ACT Project Group (http://prometheus.ms.mff.cuni.cz/act/www/),

Kiril Ribarov <ribarov@ufal.ms.mff.cuni.cz>

4. Zusammenfassung der Untersuchung

Am Ende dieses Berichts finden sich Tabellen, in denen die hier besprochenen Korpora noch einmal nach den angegebenen Kriterien zusammengefasst sind. Hier wollen wir die wichtigsten Ergebnisse kurz erläutern – dabei beziehen wir uns vor allem auf die deutschen Korpora.

Viel Material Viele ältere Texte des Deutschen sind bereits digitalisiert und öffentlich verfügbar. Diachrone Studien sind allerdings im Moment noch nicht einfach möglich, da die Unterschiede zwischen den Texten zu groß sind.

Keine Standardisierung bei Eingabe und Annotation Bisher fehlen anerkannte Standardisierungen für historische Korpora (es gibt zu viele Korpora, die sich nicht an die TEI o. ä. halten). Dies bezieht sich auf alle Punkte der Texteingabe und –vorverarbeitung: einerseits inhaltlich auf Grad der Diplomatizität, Tagsets, Annotationsrichtlinien, Annotationspraxis etc. und andererseits auf Dateiformate, Kodierung etc. Während die Unterschiede in Textauswahl und Digitalisierungsverfahren durch die Zielsetzung des Projekts begründet sind, ist nicht klar, warum so viele (insbesondere deutsche) Projekte sich nicht an korpuslinguistische Standards halten.

Unterschiedliche Abdeckung der Sprachstufen. Die verschiedenen Sprachstufen des Deutschen sind unterschiedlich gut abgedeckt. Fast alle althochdeutschen und altsächsischen Texte sind im TITUS-Korpus verfügbar, allerdings bisher meist nur nach Editionen digitalisiert. Für das Mittelhochdeutsche gibt es ein ausgewogenes annotiertes Korpus (in Trier) und einige weitere Texte, für das Frühneuhochdeutsche und neuere Texte gibt es fast nichts.

Unterschiede in den Auswertungswerkzeugen Mit der Frage des Formats der Annotation eng verknüpft ist die Frage, welches Werkzeug zur Annotation und welches Werkzeug zur Suche benutzt werden kann. Solche Werkzeuge müssen genaue Annahmen darüber machen, in welchem Format die Texte vorliegen, die sie verarbeiten sollen. Bisher gibt es kein einheitliches Suchtool.

4.1. Annotationsformate

Ein Annotationsformat legt fest, wie bibliographische und linguistische Informationen digital repräsentiert werden sollen. Sowohl die verwendeten Werkzeuge zur Annotation als auch die verwendeten Suchwerkzeuge müssen abhängig von dieser Repräsentation angepasst oder entwickelt werden. Die Annotationsstandards, die in den betrachteten Korpora angewendet wurden, sind:[7]

COCOA. Die älteste Konvention zur Annotation von Information zu einem Text ist COCOA. Sie wurde vor allem dazu verwendet, um bibliographische Informationen zu standardisieren. Die COCOA Konvention baut auf der Ascii-Kodierung auf.

[7] Neuere Standards wie der Corpus Encoding Standard CES (http://www.cs.vassar.edu/CES/), die Standards der Open Language Archives Community OLAC (http://www.language-archives.org/OLAC/metadata.html) oder IMDI (Wittenburg, Broeder & Sloman 2000) sind bisher nicht zur Anwendung gekommen.

TEI. Ein weiterer Standard der schon relativ weit verbreitet ist, ist der Standard der Text Encoding Initiative (www.tei-c.org). Dieser von Philologen und Linguisten gemeinsam entwickelte Standard ermöglicht die Annotation von bibliographischen und linguistischen Informationen auf vielen Ebenen. Zum TEI-Standard gibt es eine XML-Version.

Penn Treebank Die Korpora der Helsinki-Corpus-Familie, die morphologisch und syntaktisch annotiert sind, folgen den Annotationen der Penn Treebank (eine Baumbank mit Texten des Wall Street Journal, siehe http://www.cis.upenn.edu/~treebank/home.html) entwickelt wurden.

4.2. Werkzeuge zur Annotation

Die meisten der untersuchten Projekte haben keine Informationen zu ihren Annotationswerkzeugen veröffentlicht. Ausnahmen bilden die Korpora der Helsinki-Corpus-Familie, die die Annotationswerkzeuge der Penn Treebank verwenden und das ACT Projekt, das sehr mächtige eigene Annotationstools entwickelt hat.

4.3. Werkzeuge zur Suche

Auch bei den Suchabfragesprachen gibt es keine Einheitlichkeit, d. h., für verschiedene Korpora muss man unterschiedliche Such- und Konkordanzprogramme verwenden. Die in den hier vorgestellten Korpora verwendeten Programme sind:

PicoSearch. PicoSearch (http://www.picosearch.com/) ist ein kommerzieller Dienst der es ermöglicht im Internet verfügbare Seiten durchsuchbar zu machen. Der Corpus Scriptorum Latinorum verwendet diesen Dienst um eine Suchfunktion für seine in HTML vorliegenden Korpora bereitzustellen.

CorpusSearch. CorpusSearch (http://www.ling.upenn.edu/mideng/) ist ein Programm zur Suche in Korpora, die im Format der Penn Treebank (hier die Korpora der Helsinki Corpus Familie) annotiert wurden. Es ermöglicht die Suche nach komplexen linguistischen Strukturen (z. B. Phrasen) durch eine eigene Abfragesprache. CorpusSearch verfügt nur über ein minimales Benutzerinterface. Die Abfrage wird dem Programm in einer eigenen Textdatei, die extern erstellt werden muss, übergeben. Die Ergebnisse der Abfrage werden dann wieder in eine oder mehrere Textdateien ausgegeben. Es ist auch möglich CorpusSearch auf die Ergebnisse einer Anfrage anzuwenden.

OCP. Das Oxford Concordance Program (OCP) ist ein Werkzeug zur Erstellung von Konkordanzen, Wortlisten, Indizes. Weiter kann es einfache Textstatistiken erstellen. OCP wurde für Texte die im COCOA Format annotiert sind entwickelt. OCP ist für IBM Mainfraime Rechner und als Mini-OCP für PCs verfügbar.

WordCruncher. WordCruncher erzeugt Wortlisten, KWIC Konkordanzen und Konkordanzen von Kollokationen. Texte müssen speziell aufbereitet werden um WordCruncher mit ihnen zu verwenden.

4.4. Übersichtstabellen

Tabelle 1 Korpora und Textauswahlkriterien

Korpus	Sprache(n) / Dialekte	Zeit-raum	Textsorte	Sonstiges zu Auswahlkriterien
Bonner Frühneuhochdeutsch Korpus	Frühneuhoch-deutsch	1350-1700	verschiedene	–
Bochumer Mittelhochdeutsch Korpus	Mittelhoch-deutsch (Ober- / Mitteldeutsche Dialekte)	1070-1350	verschiedene	–
Mittelhochdeutsche Wörterbücher und Digitales Mittelhochdeutsches Textarchiv	Mittelhoch-deutsch	1050-1350	verschiedene	philologisch gesicherte Quellen
Thesaurus Indogermanischer Text- und Sprachmaterialien (TITUS)	alle indogermanische Sprachen	Keine Angaben	verschiedene	Abdeckung der Überlieferung
Codices Electronici Ecclesia Coloniensis (CEEC)	Deutsch, Latein	590-1800	Kodices	Texte aus der Erzbischöflichen Diözesan- und Dombibliothek Köln
Textkorpus von Thomas Gloning	Deutsch, Latein, Englisch	787-1900	verschiedene	–
Mittelhochdeutsche Begriffsdatenbank (MHDBDB)	Mittelhochdeutsch	1050-1350	verschiedene	–
Diachronic Part of the Helsinki Corpus	Old English, Middle English, Early Modern English	750-1700	verschiedene	ausgewogen bezügl. Dialekt und Genre
Brooklyn-Geneva-Amsterdam-Helsinki Parsed Corpus of Old English (Brooklyn Corpus)	Old English	8.Jh.	verschiedene	aus dem Helsinki-Corpus
York-Toronto-Helsinki Parsed Corpus of Old English Prose (YCOE)	Old English	8.Jh.	Prosa	aus dem Helsinki-Corpus
York-Toronto-Helsinki Parsed Corpus of Old English Poetry (York Poetry)	Old English	8.Jh.	Lyrik	aus dem Helsinki-Corpus
Penn-Helsinki Parsed Corpus of Middle English (PPCME2)	Middle English	1150-1500	verschiedene	aus dem Helsinki-Corpus
Lancaster Newsbook Corpus	Early Modern English	1653 - 1654	Newsbooktexte	aus den Thomason Tracts
Corpus of late 18th century prose	Early Modern English	1761-1789	Briefe	an einen best. Adressaten

Korpus	Sprache(n) / Dialekte	Zeit-raum	Textsorte	Sonstiges zu Auswahlkriterien
Corpus of Early English Correspondence	Early Modern English	1471-1800	Briefe	ausgewogen bezügl. soziolinguistischer Parameter
Lampeter Corpus of Early Modern English	Early Modern English	1640-1740	Traktate und Pamphlete	ausgewogen bezügl. inhaltlicher Kriterien
Corpus of Middle English Prose and Verse	Middle English	Keine Angaben	Prosa, Lyrik	–
Innsbrucker Computer-Archive of Machine-Readable English Texts	Englisch	Keine Angaben	Prosa, Briefe	–
Renaissance Electronic Texts (RET)	Early Modern English	16.Jh. und 17.Jh.	Dramen, Wörterbücher	literaturwissen-schaftlich
Glossarial Database of Middle English	Englisch	1387-1400	Glossare	opportunistisch
The Canterbury Tales Project	Middle English	14.Jh.	Märchen	verschiedene Textversionen
Corpus Scriptorum Latinorum (CSL)	Latein	1.Jh.v.Chr – 19.Jh.	verschiedene	–
La Banque de Texte de LASLA	Latein, Griechisch	1.Jh.v.Chr	Reden	–
Augustana	Versch. Indogermanische Sprachen	7.Jh.v.Chr - 20.Jh.	verschiedene	–
Corpus del Espanol	Spanisch	12.Jh. – 19.Jh.	verschiedene	–
Tycho Brahe Parsed Corpus of Historical Portuguese	Portugiesisch	1497 - 1894	verschiedene	–
Textes de Français Ancien (TFA)	Altfranzösisch	12.Jh. – 14.Jh.	verschiedene	–
Frantex	Französisch	16.Jh. – 20.Jh.	verschiedene	–
Annotated Corpora of Text (ACT)	Altkirchen-slawisch	1230 - 1450	Bibeltexte	–

Tabelle 2 Korpuseigenschaften

Korpus	Annotation	Größe	Verfügbarkeit
Bonner Frühneuhochdeutsch Korpus	BT-LM-	16.000 Zeichen	keine Angaben
Bochumer Mittelhochdeutsch Korpus	B--LM-	105 Texte	keine Angaben
Mittelhochdeutsche Wörterbücher und Digitales Mittelhochdeutsches Textarchiv	BT----	148 Texte	eingeschränkt online, CD-ROM
Thesaurus Indogermanischer Text- und Sprachmaterialien (TITUS)	BT----	k.A.	frei für wissenschaftliche Zwecke
Codices Electronici Ecclesia Coloniensis (CEEC)	BT----	65.701 Seiten	Online
Textkorpus von Thomas Gloning	B-----	35 Texte	online frei verfügbar
Mittelhochdeutsche Begriffsdatenbank (MHDBDB)	BT-LM-, zusätzlich Konzepte	127 Texte, 5,7 Mio Wörter	online frei verfügbar
Diachronic Part of the Helsinki Corpus	BT----	1,5 Mio Wörter	über ICAME
Brooklyn-Geneva-Amsterdam-Helsinki Parsed Corpus of Old English (Brooklyn Corpus)	BT-LMS	100000 Wörter	über ICAME
York-Toronto-Helsinki Parsed Corpus of Old English Prose (YCOE)	BT-LMS	1,5 Mio Wörter	über ICAME
York-Toronto-Helsinki Parsed Corpus of Old English Poetry (York Poetry)	BT-LMS	70000 Wörter	über ICAME
Penn-Helsinki Parsed Corpus of Middle English (PPCME2)	BT-LMS	1,3 Mio Wörter	über ICAME
Lancaster Newsbook Corpus	BT----	800000 Wörter	auf CD-ROM nach Anfrage
Corpus of late 18th century prose	BT----	300.000 Wörter	über das Oxford Text Archive
Corpus of Early English Correspondence	BT----	2,7 Mio Wörter	über ICAME
Lampeter Corpus of Early Modern English	BT----	1,1 Mio Wörter	über ICAME oder das Oxford Text Archive
Corpus of Middle English Prose and Verse	BT----	ca. 60 Texte	online frei verfügbar

Korpus	Annotation	Größe	Verfügbarkeit
Innsbrucker Computer-Archive of Machine-Readable English Texts	BT----	129 Texte	über ICAME
Renaissance Electronic Texts (RET)	BT----	ca. 10 Texte	für Forschungszwecke online verfügbar
Glossarial Database of Middle English	BT-LM-	k.A.	online frei verfügbar
The Canterbury Tales Project	B-N---	88 Text-fassungen	eingeschränkte Onlineausgabe frei verfügbar, Vollversion auf CD-ROM
Corpus Scriptorum Latinorum (CSL)	B-----	ca. 1000 Texte	keine Einheitliche Regelung
La Banque de Texte de LASLA	B--LM-	13 Texte	nach Anmeldung online verfügbar
Augustana	B-----	k.A.	online frei verfügbar
Corpus del Espanol	B--LM-	100 Mio Wörter	online frei verfügbar
Tycho Brahe Parsed Corpus of Historical Portuguese	B--LMS	1,8 Mio Wörter	nach Anmeldung online verfügbar
Textes de Français Ancien (TFA)	BT----	120 Texte, 3 Mio Wörter	für Forschungszwecke über ATILF[8]
Frantex	B--LM-	3500 Texte	gegen Gebühr verfügbar
Annotated Corpora of Text (ACT)	BT-LM--	700000 Wörter	online frei verfügbar

Schlüssel der Spalte Annotation (laut Angaben auf den jeweiligen Webseiten; die Angaben zu nicht online verfügbaren Korpora sind nicht überprüft):

B – Bibliographische Angaben (und andere Headerinformationen)

T – Textstruktur (strukturelle Annotation)

N – Normalisierung (hier nur angegeben, wenn explizit erwähnt; wir gehen davon aus, dass alle angeführten Korpora in einem gewissen Maße normalisieren)

L – Lemma

M – morphologische Angaben, inklusive Wortartangaben

S - Syntaxannotation

[8] ATILF: Analyse et Traitement Informatique de la Langue Française, siehe http://www.atilf.fr/.

5. Referenzen

Biber, Douglas (1993) Representativeness in corpus design. In: *Literary and Linguistic Computing* 8.243-257

Burch, Thomas; Fournier, Johannes; Gärtner, Kurt & Rapp, Andrea (Hrsg.) (2003) *Standards und Methoden der Volltextdigitalisierung. Beiträge des Internationalen Kolloquiums an der Universität Trier, 8./9. Oktober 2001.* Akademie der Wissenschaften und der Literatur, Mainz

Busa, Roberto (1974-1980): *Index Thomisticus. Sancti Thomae Aquinatis operum omnium indices et concordantiae in quibus verborum omnium et singulorum formae et lemmata cum suis frequentiis et contextibus variis modis referuntur, quaeque auspice Paulo VI Summo Pontifice consociata plurium opera atque electronico IBM automato usus digessit Robertus Busa*. Stuttgart

Evert, Stefan & Fitschen, Arne (2004) Textkorpora. In: Klabunde et al. (2004), 406 - 413

Hockey, Susan (2003) Digital Resources in the Humanities: Past, Present and Future: Towards a Universal Digital Library for the Humanities. In: Burch et al. (2003), 51-69

Kennedy, Graeme (1998) *An Introduction to Corpus Linguistics*. Longman, London

Klabunde, Ralf et al. (2004) *Computerlinguistik und Sprachtechnologie. Eine Einführung.* 2. überarbeitete und erweiterte Auflage. Spektrum Akademischer Verlag, Heidelberg

Manning, Chris & Schütze, Hinrich (1999) *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA

McEnery, Tony & Wilson, Andrew (1998) *Corpus Linguistics*. Edinburgh University Press, Edinburgh

Rissanen, Matti; Kytö, Merja; Pallander, Minna (1993) *Early English in the Computer Age: Explorations through the Helsinki Corpus* Mouton de Gruyter, Berlin

Wittenburg, P.; Broeder, D.; and Sloman, B. (2000) EAGLES/ISLE: A Proposal for a Meta Description Standard for Language Resources, White Paper. LREC 2000 Workshop, Athens. http://www.mpi.nl/world/ISLE/documents/papers/white_paper_11.pdf

Zampolli, Antonio (2004) Past & On-Going Trends in Computational Linguistics: a View from the Instituto die Linguistica Computizionale. In: *The ELRA Newsletter*, Vol 8(3), 6-16. ELRA-Paris